(6 pages)

Reg. No. : .....

Code No. : 8068 **Sub. Code : ZCAM 31**

M.C.A. (CBCS) DEGREE EXAMINATION, NOVEMBER 2023.

Third Semester

Computer Applications – Core

DATA SCIENCE AND ANALYTICS

(For those who joined in July 2021 - 2022)

Time : Third hours Maximum : 75 marks

PART A — (10 × 1 = 10 marks)

Answer ALL the questions.

Choose the correct answer :

1. Which of the following is not a part of the data science process?

   (a) Communication building

   (b) Operationalize

   (c) Model planning

   (d) Discovery

2. Data science is the process of diverse set of data through?

   (a) Organizing data

   (b) Processing data

   (c) Analysing data

   (d) All of the above

3. A correct way to preprocess the data when performing regression or classification is

   (a) Normalize the data → PCA → normalize PCA output → training

   (b) Normalize the data → PCA → training normalize → PCA output

   (c) PCA → normalize PCA output → training

   (d) Normalize the data → PCA → training

4. Which of the following is required by K-means clustering?

   (a) Defined distance metric

   (b) Number of clusters

   (c) Initial guess as to cluster centroids

   (d) All of the mentioned

5. _____ is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.

(a) Hive (b) MapReduce

(c) Pig (d) Lucene

6. *Who created Hadoop?*

(a) Doug cutting (b) JetBrains

(c) Graydon Hoare (d) Robert Griesemer

7. InfoSphere Data Stage has _____ levels of Parallelism

(a) 4 (b) 3

(c) 1 d) 2

8. InfoSphere _____ provides you with the ability to flexibly meet your unique information integration requirements

(a) Data Server (b) Information Server

(c) Info Server (d) All of the mentioned

9. _____ are at the heart of the Streams analytical engine.

(a) Operators (b) Server

(c) Client (d) File

10. Which of the following genres does Hadoop produce?

(a) Distributed file system

(b) JAX-RS

(c) Java Message Service

(d) Relational database Management System

PART B — (5 × 5 = 25 marks)

Answer ALL questions, choosing either (a) or (b).
Each answer should not exceed 250 words.

11. (a) Define data science. Why we need data science?

Or

(b) Reframe the local regression in Data science.

12. (a) Write an overview of any two unsupervised learning methods.

Or

(b) Originate K-means clustering. Specify some applications.

13. (a) Define Bigdata, Specify the characteristics of Bigdata.

Or

(b) Discuss the Hadoop different components

14. (a) How to install Infosphere BigInsights. Mention the components included in BigInsights 1.2.

Or

(b) Articulate the adaptive Mapreduce Infosphere BigInsights.

15. (a) Examine the basics of Infosphere Stream

Or

(b) Discuss the operators of Streams Processing Language.

PART C — (5 × 8 = 40 marks)

Answer ALL questions, choosing either (a) or (b)
Each answer should not exceed 600 words.

16. (a) Intervene the prerequisite probability concepts for Bayes rule.

Or

(b) Formulate the polynomial regression using R.

17. (a) Elucidate on Naïve Bayes classifier.

Or

(b) Show the Kernel density estimation in unsupervised learning

18. (a) Paraphrase

(i) Importance of Bigdata.

(ii) Bigdata use cases.

Or

(b) Interpret the application development in Hadoop.

19. (a) Analyze administrative tooling benefits in Infosphere BigInsights.

Or

(b) Generalize the Data Discovery and Visualization.

20. (a) Elucidate on industry use cases for InfoSphere Streams

Or

(b) Explain is detail the Infosphere streams tool kits.

———————